# Making Data, Making Reality

Power, Visibility, and the Production of Datasets for ML

Milagros Miceli
Weizenbaum Institute for the
Networked Society
Technische Universität Berlin
m.miceli@tu-berlin.de

## INTRODUCTION

Data is fundamental to machine learning (ML) and, more broadly, to contemporary knowledge production. Data is never raw. It is formed in the interaction with humans that collect, interpret, and classify it [4]. Working with data involves, apart from mastery of formal analysis techniques, situated knowledge and discretionary decision-making. This way, datasets used to train and evaluate ML models encode the values, prejudices, and interest of those actors involved in their creation. Furthermore, datasets hold power to render visible what they contain, and invisible what they exclude.

The present research project investigates the relationship between practices of interpretation, classification, and labeling of data and the power to shape reality through (in)visibility in datasets.

## STUDY AND INITIAL FINDINGS

With the focus set on the industrial settings, this research project explores work practices related to the creation of datasets for ML learning products. The investigation is guided by three research questions: (i) How do workers make sense of the data that will fuel ML products? (ii) What structures, standards, and organizational routines shape classification practices related to the sensemaking of data? (iii) Who decides what datasets contain and what they exclude? The project's focus has so far been directed towards the work of data annotators. A qualitative study was conducted guided by the constructivist variation [2] of grounded theory methodology. The study included several weeks of fieldwork at two annotation companies and 24 interviews with annotators, managers, and ML practitioners.

The initial findings show that annotators are subject to persistent power imbalances within their organizations, and between those and the clients commissioning the annotations. These power imbalances constrain the room for annotators' subjectivities, instrumentalizing workers to annotate data according to classifications imposed on them by other actors above their station. Power asymmetries not only manifest in labor conditions but have a definitory effect on the datasets that are produced.

Deciding what is included in and excluded from datasets is a question of power. Acts of classification are attempts to impose specific readings of the social world over other possible interpretations. Classifications are not merely a matter of sorting or describing social reality but a way of making reality by inclusion and exclusion[1]. Power, in this context, relates to the authority to lend legitimacy to certain classifications, while delegitimizing others. Analyzing the underlying assumptions and naturalized classifications involved in the creation of data and embedded in datasets means discussing the power dynamics implicit in the interpretation of data and asking who gets to decide what (and whose) data is to be included and how that data is to be interpreted. These factors decisively shape datasets and will, in time, show their effects on systems and outputs with consequences for individual identities and societal chances [3].

## CONCLUSION

Human discretion and corporate priorities intervene between data and analysis, crucially shaping data and systems, and, in some cases, the truth claims associated with systems' outputs. Assigning meaning to data is often presented as a technical matter. This investigation shows it is, in fact, an exercise of power with multiple implications for individuals and society. By examining the structures and conditions involved in the taxonomical sense-making of data, this research project offers insights into the ways machine learning can reinforce social inequities and introduces a power-aware perspective for the analysis of socio-technical systems, by showing that possible harms also relate to the arbitrary classifications that inform data creation.

## KEYWORDS

Dataset, Machine Learning, Data Annotation, Power.

## REFERENCES

[1] Pierre Bourdieu. 1992. *Language and Symbolic Power* (New ed.). Blackwell Publishers, Cambridge.

[2] Kathy Charmaz. 2006. *Constructing grounded theory*. Sage Publications, London ; Thousand Oaks, Calif.

[3] Marion Fourcade and Kieran Healy. 2013. Classification situations: Life-chances in the neoliberal era. *Accounting, Organizations and Society* 38, 8 (November 2013), 559–572.

[4] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, ACM Press, Glasgow, Scotland Uk, 1–15.